Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Исследовательский проект на тему

"Алгоритмический трейдинг: высокочастотные, низкочастотные и глобал макро стратегии"

Выполнил студент группы 191, 4 курса,
Ткаченко Егор Олегович

Руководитель КР:
научный сотрудник Соколов Евгений Андреевич

Консультант:
научный сотрудник Злотник Андрей Александрович, приглашенный преподаватель

Москва 2023

Аннотация

В своей дипломной работе я занимаюсь исследованием на тему высокочастотного трейдинга и разработки модели, способной прогнозировать изменения цены на финансовых рынках. Высокочастотный трейдинг — это стратегия торговли на финансовых рынках, основанная на быстром выполнении большого числа сделок за короткое время с использованием компьютерных алгоритмов и программного обеспечения. Она стала широко распространенной в последние годы благодаря развитию информационных технологий и возможности получения данных в режиме реального времени. В моей работе я проведу анализ данных с помощью статистических методов и машинного обучения, чтобы разработать эффективную модель, которая будет способна предсказывать изменения цены на основе исторических данных. Моя цель заключается в том, чтобы создать модель, которая будет полезна для трейдеров и инвесторов, повышая их вероятность успешной торговли на финансовых рынках.

Аннотация

I am researching high-frequency trading and developing a model capable of predicting price changes in financial markets. High-frequency trading is a strategy of trading in financial markets based on the rapid execution of a large number of transactions in a short time using computer algorithms and software. It has become widespread in recent years due to the development of information technology and the ability to obtain data in real time. In my paper, I will analyze data using statistical methods and machine learning to develop an effective model that will be able to predict price changes based on historical data. My goal is to create a model that will be useful for traders and investors, increasing their likelihood of successful trading in financial markets.

Ключевые слова: HFT, машинное обучение, финансовые рынки, Limited Order Book (LOB), Market order, Market making.

1 Введение

Мое исследование посвящено высокочастотному трейдингу на финансовых рынках. В настоящее время на высокочастотную торговлю приходится половина объема рынка в США. И от четверти до половины - на европейских рынках. С каждым годом этот объем продолжает расти, а финансовые хедж фонды наращивают интеллектуальный капитал в виде людей, которые пишут алгоритмические торговые стратегии и используют математические модели, на замену классическому "ручному"трейдингу.

Orderbook. Предметом исследования служит так называемая "Книга лимитных заявок". Она представляет собой срез спроса и предложения в конкретный момент времени. В ней можно наблюдать все лимитные заявки, размещенные на рынке в текущий момент времени.

Рынок имеет несколько уровней цен. На каждом уровне цен виден объем всех лимитных заявок по соответствующей цене. Книга заявок делится на две стороны: bid, ask - предложения и запросы. "Предложения"это заявки на покупку, гласящие о том, что участник рынка готов купить торгующийся инструмент в некотором объеме по фиксированной цене, которую он предлагает. "Запросы"же, это заявки на продаже.

Участник рынка выставляет на продажу инструмент по цене, которая его устраивает. В обоих случаях, если заявка не выставлена по рыночной цене, то участники вынуждены ожидать того момента, пока рыночная цена не сдвинется до уровня, на котором выставлена их заявка, а затем дождаться, когда их заявка встретит рыночную заявку другого игрока по противоположной операции. Если выставлена лимитная заявку на покупку - нужно дождаться пока мы не встретим рыночную операцию по нашей цене на продажу от другого участника.

Рыночная цена. Это цена, по которой реально торгуется наш инструмент в текущий момент времени. Она означает, что по этой цене можно почти мгновенно продать/купить инструмент, не дожидаясь смещения рыночной цены. Пример: имеем заявки на покупку по ценам: [22000, 22005, 22010] и заявки на продажу: [22020, 22025, 22030], то рыночная цена будет 22015, как середина по рынку между наилучшими предложениями на продажу и покупку.

Маркетмейкеры. Назревает вопрос: "А что делать если на рынке есть только лимитные ордера? Он же тогда встанет на месте!". Специально для этого существуют так называемые маркетмейкеры. Это лица, как правило большие компании, которые обеспечивают ликвидность на рынке за вознаграждение от брокера. Их задача в том, чтобы иметь запас денежных средств и торгующего актива для того чтобы непрерывно исполнять рыночные заявки обычных участников рынка. Их выгода состоит в том, что брокер выделяет им отрицательную комиссию за совершение операций. Таким образом маркетмейкеры зарабатывают деньги, а рынок наполняется ликвидностью, что позволяет участникам рынка комфортно торговать на нём.

Задача. Цель моего исследования состоит в том, чтобы пользуясь доступной информацией о состоянии рынка сделать модель максимально хорошо предсказывающую цену актива в краткосрочном горизонте прогнозирования (В моем случае это следующие 10 милисекунд). На основе истории книги заявок и сделок нужно сделать максимально полезные признаки, имеющие в себе полезную информацию о текущей динамике цены. Чем больше полезной информации мы имеем, тем больше преимущества у нас относительно других участников. Стоит помнить что в высокочастотном трейдинге мы не просто забираем прибыль из воздуха, ведь в это же самое время множество конкурентов занимаются тем же самым, а значит задача становится сложной.

Из-за чего происходит сдвиг цены? Рыночные заявк, при отсутсвии встречных рыночных заявок начинают истощать лимитные заявки по наилучшей цене. Когда заявки по наилучшей цене по какой-либо из сторон заканчиваются - происходит смещение к следующему уровню ордеров. И как правило после этого достаточно быстро с противоположной стороны начинают поступать лимитные заявки на уровень ближе к новой цене, так как участники рынка с той стороны понимают, что цена сдвинулась от них, а значит ожидаемое время исполнения их ордеров увеличилось.

Виды событий на рынке. Отмены. Исполнения. Я уже ввел такие термины, как поступление лимитных и рыночных заявок, а теперь напишу про оставшиеся события. Помимо выставления заявок игроки рынка также могут отменять имеющиеся у них лимитные заявки. Это нужно как раз для того, чтобы адаптироваться по изменяющуюся ситуацию на рынке. Если цена сдвинулась далеко от нас, то имеет смысл переставить свои заявки. А если до нашей лимитной заявки все же доходит очередь, то происходит исполнение нашей заявки. Изучение частот возникновения всех этих событий является очень информативным и может многое рассказать о текущей динамике рынка. В дальнейшем я убедился в этом, в одном из экспериментов.

Спотовый рынок. Рынок фьючерсов. Существует два вида торговли активами. Спотовая торговля, это когда происходит торговля непосредственно активом и он поступает на баланс в этот же момент времени. Рынок фьючерсов - это рынок одного из деривативных инструментов, когда цена на торгуемый контракт зависит от цены инструмента, на основе которого был сделан этот контракт. Есть много разных деривативов, но сейчас речь идет только о фьючерсах. Фьючерс - это контракт, который по сути является договором на куплюродажу по фиксировнной цене к фиксированному моменту времени.

Bid, Ask. Так называют стороны ордербука. Bid, т.е "ставка это ордер на покупку актива, за выставленную, фиксированную участником рынка цену. Ask, т.е "запрос ордер на продажу актива участником рынка, за цену, которую он запрашивает у потенциального покупателя.

2 Обзор литературы

good one

2.1 Fragmentation, Price Formation, and Cross-Impact in Bitcoin Markets.

Основная идея этой статьи состоит в том, что разные части рынков по своему влияют на динамику цены. Что подразумевается под разными частями рынка? Во-первых криптовалюты торгуются сразу на множестве разных бирж. Помимо бирж у нас также имеется два вида рынков - спотовый и фьючерсный. Мы давали определение этих рынков выше. Основное отличие рынка биткоина от рынка традиционных акций, облигаций, состоит в том, что он намного сильнее фрагментирован между разными биржами, а также в более высокой волатильности. Дело в том, что биткоин фрагментирован по множеству нескольких высоколиквидных бирж, внутри которых торговый объем также размывается на различные производные контракты от него. Биржи работают независимо друг от друга и подчиняются разным условиям регулирования, в зависимости от их места базирования. Это приводит к распространенному явлению арбитража между биржами.

В этой статье авторы стараются найти ответы на следующие вопросы:

- 1. На какой части рынка информация о цене появляется раньше всего?
- 2. Каким именно образом на цену на одной площадке влияет поступление информации на соседнюю?
- 3. Как сильно разница в воздействии на цену между площадками и инструментами позволяет её прогнозировать?
- 4. И наконец можно ли на основании всей этой информации создавать стратегии, способные приносить прибыль при торговле в реальном времени?

Полезные признаки из статьи. Представлены следующие признаки на основе ситуации на разных частях рынка:

Trade Imbalances. Для выбранного рынка i для каждого из временного горизонта $\delta \in \{100, 250, 500, 1000, 2000\}$ милисекунд,в текущий момент времени t имеем:

$$TFI_t^{i,\delta} = B^i_{[t-\delta,t]} - S^i_{[t-\delta,t]} \tag{1}$$

Где $B^i_{[t-\delta,t]}$ и $S^i_{[t-\delta,t]}$ - суммарные объемы покупок и продаж за выбранный временной промежуток, на конкретной части рынка соответственно. Trade Imbalance позволяет улавливать короткосрочные изменения направления потока сделок.

Past Returns. Он считается также для каждого временного интервала из вышеописанных, все последующие фичи в рамках этой статьи тоже. p_t^i это средняя цена за короткий промедуток в 50 мс. $p_t^i = \frac{1}{A}\Sigma_{(a,p)\in T_t^i}(a*p)$, где (a,p) - пары количество/цена всех сделок исполненныъ за это время. Проще говоря величину p_t^i можно воспринимать как средняя взвешенная по объемам цена актива. Затем мы используем эту величину для подсчета финальной фичи:

$$PRET_t^{i,\delta} = (\frac{p_t^i}{p_{t-\delta}^i} - 1) \tag{2}$$

Past Returns - степень изменения цены за промежуток времени δ. Это помогает выявлению лидеров и отстающих среди рынков и значимое изменение этой статистики у лидерах дает понять что отстающие тоже подхватят тренд.

Mean Divergence Для $\Delta \in \{5, 9, 19, 38, 75, 150, 300, 600\}$ секунд, и для всех пар частей рынка i, j:

$$DIV_t^{i,j,\Delta} = d(p_t^i, p_t^j) - rolling^{\Delta}(d(p_t^i, p_t^j)) \tag{3}$$

Где $d(p, q) = (\frac{p}{q} - 1)$, rolliing() - скользящее среднее по величине, за выбранный промежуток времени.

Идея такая что если есть какой-то рынок, на котором наш актив дешевле, то там ожидается скачок цены до "рыночной". Для неё посчитывается отношение цен между рынками и скользящее среднее этих отношений. И если отношение в текущий момент ощутимо отличается от этой же величины "в среднем"за промежуток времени, то это может сигнализировать о том, что в данный момент наблюдается отклонение от тренда этого отношения, а значит последует регуляция этого явления.

2.2 Mid-price prediction based on machine learning methods with technical and quantitative indicators.

Аннотация. Предсказание фондового рынка является сложной задачей, в которой методы машинного обучения смогли преуспеть за последние несколько лет. В это статье авторы вводят около 270 признаков, созданных на основе технических индикаторов, количественного анализа. Они изучали признаки разделив их на группы. Цель авторов состояла в том, чтобы используя как можно меньшее количество признаков получить наиболее хорошую модель, так как большое количество признаков далеко не всегда хорошо, а иногда может только испортить модель.

Из всего списка мной были выбраны и запрограммированы следующие признаки. Также поясню их смысл. Особенность подсчета фичей такая, что мы смотри на наш датасет как на блоки размером в 10 наблюдений. Каждый блок имеет свою цену закрытия - цену в последний момент блока. Эта блочность учитывается при подсчете фичей.

Accumulation Distribution line - смотрим на кумулятивную сумму произведения цены закрытия на объем. Позволяет видеть направление текущего тренда. Угасает он или растет.

$MoneyFlowMultiplier = \frac{C_t - L_t}{H_t - L_t}$

$MoneyFlowVolume = MoneyFlowMultiplier * BlockPeriodVolume$

$\text{ADL} = ADL_{t-1} + MoneyFlowVolume_t$

C_t, L_t, H_t - цена закрытия, минимальная, максимальная цены внутри текущего блока. BlockPeriodVolume - суммарный объем за блок.

Average Directional Index - указывает на силу и направление тренда в моменте. Позволяет понять насколько устойчива ситуация на данный момент.

$+DI = \frac{Smoothed+DM}{ATR}$, $-DI = \frac{Smoothed-DM}{ATR}$, $DX = \frac{\|+DI-DI"\|}{\|+DI-DI"\|}$,

$ADX = rollingMean(ADX_{[t-14,t]})$.

Где: $+DM = H_t - H_{t-1}$, $-DM = L_{t-1} - L_t$

Smoothed+-DM=ExponentialMovingAverage(+-DM, 1/14)

$\text{ATR} = \text{mean}(\text{TrueRange}_{[t-14:t]})$

$\text{TrueRange} = \max(\text{H}_t - L_t, |H_t - CL_{t-1}|, |L_t - CL_{t-1}|)$

Change Momentum oscillator - указывает насколько сильно изменилась цена закрытия за выбранный период. Позволяет замечать развороты тренда или ноборот его усиления.

$$S_u = \sum_{i=1}^{19} CL_i * 1_{CL_t > CL_{t-19}} \tag{4}$$

$$S_d = \sum_{i=1}^{19} CL_i * 1_{CL_t < CL_{t-19}} \tag{5}$$

$$CMO = \frac{S_u - S_d}{S_u + S_d} \tag{6}$$

Momentum - указывает на скорость изменения цены за выбранный период. Позволяет выявлять разворот тренда или его продолжение.

$$MOM = CL - CL_{t-1} \tag{7}$$

Rate of Change - скорость с которой изменяется цена за выбранный период.

$$ROC = CL_t - \frac{CL_t - CL_{t-12}}{CL_{t-12}} \tag{8}$$

Stochastic Relative Strength Index - получается на основе индикатора RSI. Указывает на перекупленность либо недооцененность актива на текущий момент. Позволяет понять ожидать ли снижения или повышения цены в ближайшее время.

$$Stoch_{RSI} = \frac{RSI_{curr} - RSI_{Low_{10}}}{RSI_{High_{10}} - RSI_{Low_{10}}} \tag{9}$$

$$RSI = 1 - \frac{1}{(1 + Relative_{Strength})} \tag{10}$$

$$Relative_{Strength} = \frac{AG_{14}}{AL_{14}} \tag{11}$$

$$AG_{14} = \sum_{i=1}^{14} CL_{d_t} 1_{CL_{d_t} > CL_{d_{t-1}}} \tag{12}$$

$$AL_{14} = \sum_{i=1}^{14} CL_{d_t} 1_{CL_{d_t} < CL_{d_{t-1}}} \tag{13}$$

$$CL_{d_t} = CL_t - CL_{t-1} \tag{14}$$

Realized Volatility - измеряем волатильность внутри выбранного периода. У нас может быть низкая волатильность за месяц, при этом очень ощутимая внутридневная, которая бы не была учтена, если бы мы считали волатильность как стандартное отклонение по дневным ценам закрытия за месяц.

$$RV = \sum_{i=1}^{N} r_t^2, \text{ где } r_t = \log P_t - \log P_{t-1} \tag{15}$$

Realized Kernel - Введене ядровой функции позволяет нам более точно оценивать как предыдущие движения цены влияют на её текущее формирование. Ядровая функция позволяет улавливать сложные нелинейные зависимости, в отличии от других более простых методов.

Сама формула:

$$K(X) = \sum_{h=-H}^{H} k(\frac{h}{H+1})\gamma_h, \ \gamma_h = \sum_{j=|h|+1}^{h} x_j x_{j-|h|} \tag{16}$$

Где $k(x)$ ядровая функция весов.

Существует множество вариантов выбора ядровой функции для такой задачи, но мы остановимся на следующей: $k(x) = 1 - 3x^2 + 2x^3$

Jump Variation - как я понял этот индикатор нужен для отслеживания моментов когда рынок наиболее волатилен и непредсказуем. Смотрим на отношение модуля ценового разрыва к размеру окна внутри которого смотрим на эту величину. Интуитивно кажется, что полезнее будет брать средние и

маленькие размеры окна.

$$JV(X)_t = max(RV(X)_t - BV(X)_t, 0) \tag{17}$$

$$RV(X)_t = \Sigma_{i=1}^{n}(r(X)_i)^2, \tag{18}$$

$$BV(X)_t = \frac{\pi}{2}\Sigma_{i=2}^{n}|r(X)_i||r(X)_{i-1}|, \tag{19}$$

$$r(X)_i = X_i - X_{i-1} \tag{20}$$

2.3 Stochastic Market Microstructure Models of Limit Order Books

Это совместный доклад Rama Cont и Costis Maglaras. Columbia University. по одноименной статье, которой нет в открытом доступе. В нём говорится о микроструктуре электронных финансовых рынков на основе лимитных ордербуков и ньюансах, которые стоит учитывать при анализе рынков и построении стратегий. Общая схема такая: существует множество портфолио менеджеров в управляющих фондах, которые принимают решения по сделкам. Далее они направляют ордера в свой алгоритмический торговую систему, которая занимается исполнением этого решения. Эта система распределяет большой ордер среди множество бирж и даркпулов.

Даркпулы это площадки на которых покупатели и продавцы могут торговать активами без публичного раскрытия информации о сделках. Это выгодно участникам даркпула, когда информация о сделке может предоставить сигнал другим участникам рынка. Даркпулы существуют для крупных инвесторов и фондов, которые торгуют очень большими объемами активов, и которые не хотят чтобы их торговые стратегии были раскрыты, а также не хотят вызывать значительные колебания цены, которые могут быть вызваны влиянием этих сделок на открытые биржи. Это является очень важным для крупных институциональных инвесторов - не дать другим скромпроментировать нашу стратегию. Для этого крупные ордера исполняются не сразу, а равными частями в течении дня, в моменты, когда это оптимально делать.

Как уже было сказано ранее: задача маркетмейкеров в этой системе поставлять ликвидность на рынок. Они не держат никаких долгосрочных позиций и стараются всегда поддерживать баланс между активами и свободными средствами. Их цель это моделировать динамику цены в короткосрочном периоде. Маркетмейкеры подвержены рискам из-за сильных скачков цены.

Алгоритмическую торговлю можно разложить на следующие составляющие:

1 **Предсказание внутридневных переменных ранка:** объем, спред, волатильность, глубина рынка. Очень важно учитывать текущее настроение рынка, отличие его динамики от предыдущей, насколько он сейчас спокойный, агрессивный, какой имеется тренд и так далее.

2 **Короткосрочные сигналы движения рынка.** Как правило на основе статистических методов мы моделируем динамику поведения рынка и его переменных, для того чтобы правильно оценивать текущий момент.

3 **Установление расписания выставления ордеров:** Мы делим наш ордер на кусочки и решаем когда именно и с какой частотой исполнять эти кусочки.

4 **Оптимальное исполнение кусочков большого ордера** Сколько будет стоить исполнить этот кусок? Выгодно ли это сделать в текущий момент?

5 **Распределение крупных ордеров** На какой бирже это будет выгоднее всего прямо сейчас?

На рисунке изображены объемы торгов в течении дня. Примечательно, что торговый день начинается с небольшого пика, так как уже успели выйти какие-то новости и участники рынка хотят исполнить новые сделки. Также в конце дня виден сильный всплеск по активности, так как множество трейдеров хотят закрыть свои позиции, или наоборот укрепить их.

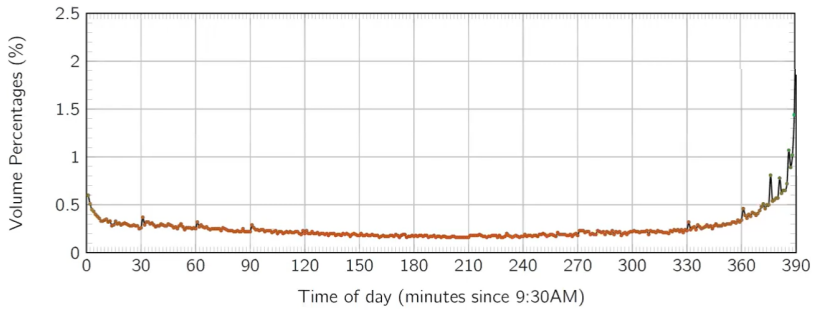

Рис. 2.1: Иллюстрация внутридневного объема торгов.

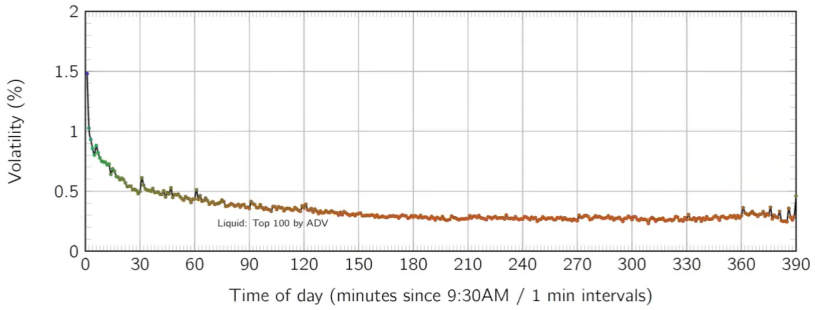

Figure: Cross-sectional avg intraday volatility profile for US equities, S&P100

Рис. 2.2: Иллюстрация внутридневной волатильности цены.

На втором рисунке изображена внутридневная волатильность. Когда торговый день открывается актив ведет себя очень изменчиво, опять таки из-за разных событий, которые успели произойти за время закрытия рынка. К концу дня она становится меньше и меньше.

Далее говорится о том что с каждый ценовой уровень с обеих сторон имеет свою частоту поступления ордеров именно на этот уровень. Также у нас есть изменяеющаяся частота отмен ордеров и исполнений. Если мы хотим выставить большой ордер эти переменные помогут нам в оценке времени на исполнение этого одерда, что бывает крайне полезно.

Надо пытаться учитывать особенности рынка, время, потоки выставления/отмены/исполнения ордеров. В разное время эти соотношения отличаются и нужно это учитывать при предсказании. В течении дня у нас наблюдается разная волатильность, разный спред, У аска и бида потоки "ивентов"очевидным образом отлчиаются, тем более на разных уровнях эти потоки также разные, в зависимости от ситуации.

Можно моделировать распределением Пуассона поступления ордеров и их выполнения (я немного пытался в это, но решил скипнуть и вернуться потом). Эти частоты зависят от уровня цены, расстояния от best prices, других частот, и так далее.

Частота изменения цены. При выставлении ордера надо учитывать успеет ли он исполниться, или какая часть от него скорее всего исполнится. Другие игроки тоже выставляют ордера и это надо учитывать. Хотим моделировать задержку. Частота отмен ордеров зависит от состояния ордербука. Большая часть ордеров отменяется из малых стопок. Другие частоты также повышаются когда имеем малые queues. Далее приводят пример простой модели для оценки вышеперечисленных величин.

Во второй части доклада Рама Конт рассказывает о упрощенной модели ордербука, где фигурируют только наилучшие bid/ask цены и соответствующие объемы. Далее он вводит две функции распределения $f(x, y)$ и $\tilde{f}(x, y)$. Каждая из них выдает вероятность того, что после повышенияонижения цены объемы лучших стопок будут x и y соответсвенно. Далее он вводит net order flow process, которые задействует еще и евенты и получаем динамику ордербука после всех ивентов. Далее он ещё подробнее погружается в эту концепцию. Я записал для себя его идею на будущее, чтобы на текущий момент заниматься извлечением информации из менее сложных моделей и статистик, но очень заинтересовался и обязательно опробую в последующей работе над моделью.

Далее я выделил самые на мой взгляд важные и полезные для моей задачи концепции из доклада:

- Моделирование частот исполнения, отмены, поступления ордеров
- Оценка различий текущего состояния переменных рынка от внутридневного среднего.
- Моделирование времени полного исполнения best bid/ best ask ордеров.

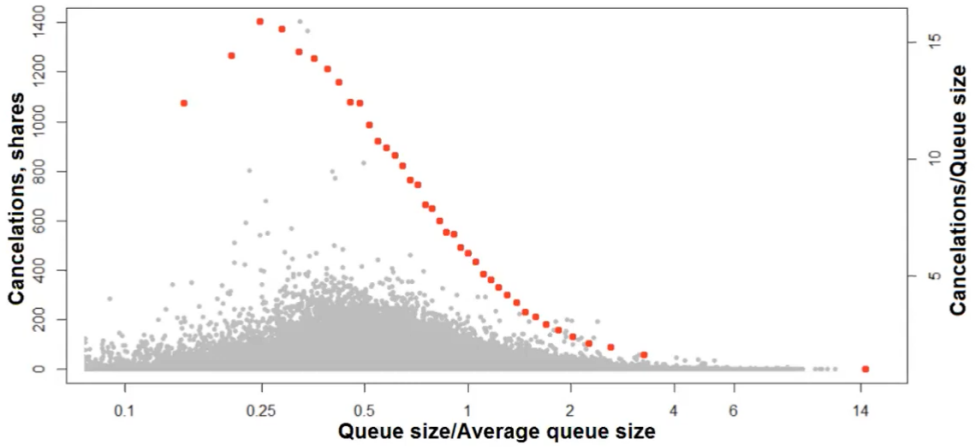

Рис. 2.3: Зависимость количества отмен ордеров от размера очереди ордеров.

- Чем лучше мы понимаем динамику отмены ордеров, тем выше точность оценки задержки перед исполнением выставленных ордеров. Нужны отдельные модели для оценки этого эффекта.

- Частота отмен непостоянна и зависит напрямую от состояния ордербука.

В своей работе на основе этих идей я разработал свои собственные признаки, которые оказались очень даже полезными. О них я расскажу в другом разделе.

3 Главы. Описание методов. Теоретический анализ. Эксперементальные исследовани.

3.1 Метрика.

3.1.1 Формула

Допустим $\mathbf{y}_{\text{true}}$ и $\mathbf{y}_{\text{pred}}$ это векторы предсказанных классов и настоящих значений, соответсвенно. Пусть $\mathbf{C}$ Это матрица ошибок такая, что c_{ij} - это количество элементов относящиеся к классу i, которые предсказывающая модель определила к классу j, посчитанная на $\mathbf{y}_{\text{true}}$ и $\mathbf{y}_{\text{pred}}$. Пусть $\mathbf{W}$ матрица весов определенная следующим образом:

$$\mathbf{W} = \begin{bmatrix} 1.9 & 0 & -2 \\ -0.3 & 0 & -0.3 \\ -2 & 0 & 1.9 \end{bmatrix}$$

Почему так?

Затем введём матрицу соответсвий $\mathbf{H}$ как поэлементное произведение матриц $\mathbf{C}$ и $\mathbf{W}$:

$$\mathbf{H} = \mathbf{C} \odot \mathbf{W}$$

Здесь, $\odot$ означает поэлементное произведение матриц.

hit_matrix_sum: сумма элементов матрицы соответсвий $\mathbf{H}$:

$$\text{hit\_matrix\_sum} = \sum_{i,j} H_{i,j}$$

action_count: это сумма первого и третьего столбцов матрицы $\mathbf{C}$:

$$\text{action\_count} = \sum_{i} C_{i,1} + \sum_{i} C_{i,3}$$

В итоге наша матрика вычисляется, как:

$$\text{metric}(\mathbf{y}_{\text{true}}, \mathbf{y}_{\text{pred}}) = \begin{cases} 0, & \text{Если action\_count} = 0 \\ \frac{\text{hit\_matrix\_sum}}{\sqrt{\text{action\_count}}}, & \text{Иначе} \end{cases}$$

3.1.2 Обоснование метрики.

Почему мы используем именно эту метрику, а не общепринятые "классические"по типу accuracy, f1-score? Если обратить внимание, то наша метрика учитывает роль каждого класса в прогнозировании. Она не просто штрафует нас если мы не угадали класс, и награждает если угадали. Если посмотреть на матрицу весов, то видно, что штраф если цена изменилась, но не в ту сторону куда мы спрогнозировали - сравнительно больше, чем если цена на самом деле не изменилась.

Это имеет очевидную реальную интерпретацию. Если мы вошли в сделку, а цена пошла в противоположную сторону, то наши убытки со сделки будут больше, чем если цена не изменилась в следующий момент, так как первый случай подразумевает более рисковое положение и сигнал, что из сделки стоит выйти как можно скорее, в то время как, если цена не изменилась, мы можем не выходить из сделки, а дождаться более благоприятного момента (если конечно позволяет наша торговая стратегия).

3.1.3 Оптимальное значение и нижняя граница.

Метрика это, конечно, хорошо, но на первый взгляд не очень ясно, как её интрепретировать. Мы собрали признаки, обучили модель, подобрали гиперпараметры, посчитали метрику. Хорошая она или плохая? Если хорошая, то насколько далека от идеала?

Метод	Трейн	Валидация	Тест
Идеальная метрика	5900.452	3779.450	3643.378
Все нули	0.0	0.0	0.0
Вероятностный подход	-608.577	-376.399	-398.249

Таблица 3.1: Бенчмарки метрики

3.2 Данные

В качестве наблюдаемого актива была выбрана криптовалюта Биткоин из-за своей новизны и неизученности относительно традиционных финансовых активов, а также повышенной внутридневной волатильности. Данные имеют следующую структуру: book - высокочастотные наблюдения ордербуков. trades - сделки на аккумулированные покупкуродажу с определенной частотой ticker - котировки. В отличии от наблюдения ордербука имеют более высокую частоту и содержит в себе лишь информацию об best ask/bid объемах и ценах.

Аналогичные данные для спотового рынка: bookSpot, tradesSpot, tickerSpot. Вышеупомянутые - для фьючерсных контрактов на биткоин.

target - цена в последующие 10 милисекунд.

Размер данных: 33 млн строк. Далее данные для тренировочной, валидационной и тестовой выборок были разбиты по пропорциям 2.5/1/1.

На основе доклада *Stochastic Market Microstructure Models of Limit Order Books* я выделил следующие признаки, которые могут быть полезны при построении модели. Далее я опишу сами признаки и поясню логику, по которой они должны действовать.

Для каждого $h \in \{500ms, 1500ms, 2500ms, 5000ms\}$:

1.$AskExecuteRate_h$, $BidExecuteRate_h$ - частота исполнения ордеров на стороне ask, bid, посчитанная на основе данных за последние δ милисекунд.

2.$TimeBidExecute_h$,$TimeAskExecute_h$ - сколько времени потребуется для полного исполнения ордеров по наилучшему предложению ask/bid, если частота исполнения ордеров посчитана на основе последних δ милисекунд.

$$TimeExecute^a = \frac{q^a_{best}}{ExecuteRate^a_h},\ TimeExecute^b = \frac{q^b_{best}}{ExecuteRate^b_h} \tag{21}$$

3.$TimeSincePriceChange$ - сколько времени прошло с последнего изменения цены вверх либо вниз.

4.$Std_{h_{Vol}}$ - стандартное отклонение объема за выбранный временной промежуток.

5. BuyAmount / SellAmount/ TotalAmount - сколько ордеров было исполнено между текущим и предыдущим снепшотом ордебука

6. $RelVol_h$ - отношение объема ордербука в текущий момент к среднему объему ордербука за выбранный временной период

7. $Hour$ - текущее время

8.$PartOfDay$ - текущая часть дня (ночь/утро/день/вечер)

9. $NewAskDiff$ / $NewBidDiff$ - просто дифф best ask/ best bid со значением в предыдущий снепшот ордербука. С поправкой на то, что когда изменяется цена, то текущий best ask/bid это по сути тот, что стоял на уровень глубже наилучшего в предыдущий момент времени. Так как изменение цены случается тогда, когда объема best bid/ask не хватает для исполнения всех рыночных ордеров, и уровень сдвигается к следующим предложениям.

Мотивация за фичами:

1.$AskExecuteRate_h$, $BidExecuteRate_h$ - частота исполнения ордеров должна давать представление о том, насколько сейчас активный рынок, как много исполняется сделок.

2.$TimeBidExecute_h$,$TimeAskExecute_h$ - чем меньше это значение тем больше шанс того, что наилучшая цена истощится и произойдет сдвиг к следующей стопке ордеров.

3.$TimeSincePriceChange$ - если с последнего изменения цены прошло много времени, то велик шанс, что это изменение должно вот-вот произойти.

4.Std_h^{Vol} - смотрим насколько изменчив объем на рынке в последнее время. Разные горизонты позволяют заметить разницу с более большими по времени окнами.

5. BuyAmount / SellAmount/ TotalAmount - опять таки позволяет судить о том, насколько активна каждая из сторон и рыночную активность в совокупности.

6. $RelVol_h$ - позволяет уловить "перегруженность"или "недогруженность"ордербу по количеству заявок. Кажется что если их резко стало становиться больше обычного, то это может сказать о том, что либо другие игроки начинают понимать куда пойдет цена и раскидывают выгодные лимитки, либо движение по наилучшим ценам замедлилось и ордеры начинают скапливаться (но тогда их должны будут отменять, так что не уверен во второй половине сказанного).

7. $Hour$ - в видео демонстрируется то, что в разное время рынок ведет себя по-разному. Так что будет полезным подавать нашей модели информацию сколько времени прошло с открытия.

8.$PartOfDay$ - то же самое, но более обобщённая фича.

9. $NewAskDiff$ / $NewBidDiff$ - возможно получится в моменты разрыва понимать насколько сильно рынок протолкнул ордера на втором уровне.

3.3 Эксперименты

Я рассматривал две модели: Линейную регрессию и метод CatBoost, основанный на градиентном бустинге. Первый я использовал в качестве базовой модели, чтобы опять-таки сравнивать дальнейшие результаты с результатом линейной регрессии, для понимания прогресса.

3.3.1 Перебор гиперпараметров

Модель Catboost по умолчанию работала плохо. Она сильно переобучалась на тренировочных данных. Поэтому подбор гиперпараметров я начал с параметров, отвечающих за регуляризацию модели. Первым шагом был случайный поиск по сетке гиперпараметров, для того чтобы за эффективное время найти хорошее начальное приближение оптимальных параметров, а далее продолжать их подбор болеее аккуратно.

Параметр	**Значения**
Learning rate	$\mathcal{U}(0.01, 0.29)$
Depth	$\text{RandInt}(6, 17)$
L2 leaf regularization	$\mathcal{U}(10^{-6}, 10)$
Bagging temperature	$\mathcal{U}(0, 10)$
Border count	$\text{RandInt}(32, 256)$
Iterations	$\text{RandInt}(100, 300)$

Таблица 3.2: Гиперпараметры

К сожалению такой перебор не дал мне сразу хорошие результаты, так что пришлось более аккуратно подходить к выбору параметров.

3.3.2 Общие поиски на больших итерациях.

subsample	random strength	max depth	l2 leaf reg	iterations	Train	Valid	Test
0.725	3	6	10	150	538	200	70
0.725	3	6	10	100	432	125	135
0.725	2	7	10	125	558	136	141
0.7	2	7	10	125	**566**	**135**	**150**
0.7	2	8	10	125	659	134	142
0.7	2	8	10	100	516	124	143

Таблица 3.3: Значения параметров и результаты модели

По итогу я решил проверить, вдруг причиной переобучения является слишком большое количество деревьев в алгоритме и оказался прав. После того как я выбрал 20 итераций вместо 100-150, качество не стало выше на валидации и тесте, зато по метрике на трейне было видно что алгоритм стал намного меньше переобучаться, значит был сделан шаг в верную сторону.

subsample	random strength	max depth	l2 leaf reg	iterations	Train	Valid	Test
0.75	1	7	10	20	310	116	117
0.7	1	7	10	20	317	101	113

Таблица 3.4: Значения параметров и результаты модели при 20 итерациях

3.3.3 Class weights

Затем я применил к алгоритму параметр, которые учитывает вес классов на трейне, что сильно повляло на метрику. При фиксированных параметра из последней таблицы я рассмотрел разные веса коэффициентов.

"1"weight	"0"weight	1"weight	Train	Valid	Test
default	10	20	310	116	117
default+10%	default	default	399	185	164
default+5%	default	default+10%	386	197	168
default+10%	default	default+10%	390	194	167
default+10%	default-10%	default+10%	418	185	168
default+10%	default%	default+10%	416	188	172

Таблица 3.5: Значения параметров и результаты модели при 20 итерациях

3.3.4 Learning rate

Далее я перебирал значения learning rate [0.05, 0.1, 0.33, 0.5, 0.13, 0.033, 0.01,0.001, 0.0001]

И самое лучшее значение lr было равным 0.0001 и **наилучшие значения метрики получились: 400, 203, 181** на трейне, валидации и тесте соотсветсвенно. Все прочие параметры не менялись с вышеупомянутых.

3.3.5 Постпроцессинг.

Важной частью построения прогнозов является постобработка предсказаний модели. Это связано с тем, что модель возвращает вероятности каждого класса. Без постобработки прогнозируемым классом выбирается тот, у которого наибольшая вероятность среди всех вероятностей. А это означает, что в случаях, когда модель имеет низкую уверенность в прогнозе, мы все равно спрогнозируем движение цены, а значит с очень большим шансом не угадаем и испротим метрику. Я решил подобрать для позитивного и негативного классов пороги, начиная с которых мы принимаем прогноз как уверенный. Если вероятности обоих классов ниже порогов, то присуждаем класс "0 чтобы застраховать себя от рисков. Это я собираюсь сделать при доработке модели в будущем.

3.4 Дальнейшие планы

В дальнейшем хотелось бы изучить больше материалов для того чтобы добавить больше фичей в модель. Что касается модели хочется дальше еще лучше подобрать параметры, а также применить такую технику как ансамблирование моделей, для повышения качества и надежности. Прочитать статьи по вероятностному моделированию процессов на рынке для того чтобы с их помощью создавать новые фичи.

4 Заключение

Моя работа показала, что разработка модели прогнозирования цен с использованием статистических методов и машинного обучения может быть эффективным подходом к улучшению результатов высокочастотного трейдинга. Есть куда расти и множество подходов которые предстоит испробовать. Я не собираюсь останавлливаться на достигнутых резульатах, а продолжу работать над алгоритом, чтобы повысить его качество в разы. В идеале хочется добиться того, чтобы мой алгоритм мог торговать в плюс на реальных данных.

Итоговый код доступен по ссылке:
https://github.com/HalcyonForest/HFT_diploma

Список литературы (или источников)

1. Mid-price prediction based on machine learning methods with technical and quantitative indicators. Adamantios NtakarisI, Juho Kanniainen, Moncef Gabbouj, Alexandros Iosifidis.

2. Fragmentation, Price Formation, and Cross-Impact in Bitcoin Markets. Jakob Albers. Department of Statistics, University of Oxford Oxford, UK

3. Feature Engineering for Mid-Price Prediction With Deep Learning. Adamantios Ntakaris, Giorgio Mirone, Juho Kanniainen

4. Stochastic Market Microstructure Models of Limit Order Books. Costis Maglaras, Columbia University; Rama Cont, University of Oxford. *https://urlis.net/hd4b7fex*

5. M. Avellaneda and S. Stoikov, “High-frequency trading in a limit order book,” Quantitative Finance, vol. 8, no. 3, pp. 217–224, Apr. 2008

6. F. Guilbaud and H. Pham, "Optimal High Frequency Trading with Limit and Market Orders," SSRN Electronic Journal, 2011

7. R. Cont and A. de Larrard, "Price Dynamics in a Markovian Limit Order Market," SSRN Electronic Journal, 2012

8. T. Ho and H. R. Stoll, "Optimal dealer pricing under transactions and return uncertainty," Journal of Financial Economics, vol. 9, no. 1, pp. 47–73, Mar. 1981

9. W. Kenton. "Order Book," Investopedia. https://www.investopedia.com/terms/o/order-book.asp